TACO: Towards Task-Consistent Open-Vocabulary Adaptation in Video Recognition

Minghao Zhu **Xiao Lin** **Mengxian Hu** **Xun Zhou**
Liuyi Wang **Xiaoyan Qi** **Chengju Liu*** **Qijun Chen**
Tongji University, Shanghai, China
{zmhh_h, linx_xx, humengxian, zhouxun,wly
2331853, liuchengju, qjchen}@tongji.edu.cn

Abstract

Adapting CLIP for open-vocabulary video recognition necessitates a delicate balance between newly acquired video knowledge and the pretrained generalization. While existing studies pursue this generalization-specialization trade-off with additional regularizations or constraints, we argue that they overlook the deviation of representations beyond the fine-tuning data distribution, resulting in suboptimal adaptation effects. We believe such deviation is inherited from the inconsistency between the fine-tuning and evaluation objectives, where model optimization is restricted to the known training distribution but evaluated on unseen ones. In this paper, we introduce *TACO*, a simple yet effective framework to mitigate the potential negative effects induced by this inconsistency. Our key insight is that adaptation should preserve OOD-relevant alignment beyond the training distribution. To this end, we propose *Relative Structure Distillation*, which regularizes the relative geometry of the representation space and suppresses harmful alignment shift during training. We further decouple the representation space from the optimization space with a lightweight specialization projection, allowing task-specific adaptation without directly overspecializing the representations used at test time. *TACO* establishes state-of-the-art performance on diverse benchmarks under cross-dataset and base-to-novel settings. Code will be released at https://github.com/ZMHH-H/TACO.

1 Introduction

Open-vocabulary learning aims to identify novel visual concepts specified by language vocabularies unseen during training phase. This setting has been propelled by the rapid progress of large-scale Vision–Language Models (e.g., CLIP [37], ALIGN [17], Florence [50]), which learn aligned multimodal representations from massive image-text pretraining. With their strong generalization to unseen visual concepts, these foundation models are highly practical in real-world applications and have been widely adapted to various downstream tasks [6, 11, 28, 51, 54]. Inspired by this progress, recent studies have explored adapting CLIP for general video recognition [33, 38, 55], avoiding the substantial cost of collecting large-scale video-text data and pretraining from scratch.

While CLIP provides a strong starting point for visual understanding, effective adaptation for open-vocabulary video recognition remains challenging - specifically in empowering models to accurately capture the temporal dynamics encoded in videos of unknown categories. The key challenge is to introduce video-specific knowledge while preserving the pretrained generalization, a problem commonly framed as the generalization/specialization trade-off. To strike this balance, recent studies (e.g., Open-VCLIP [42], FROSTER [15]) typically seek a middle ground between the two aspects by incorporating carefully designed regularizations during fine-tuning. Despite their promising results, we argue that these methods overlook representation variation beyond the training distribution, as their optimization objectives are strictly confined to the fine-tuning data distribution (i.e., ID space).

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

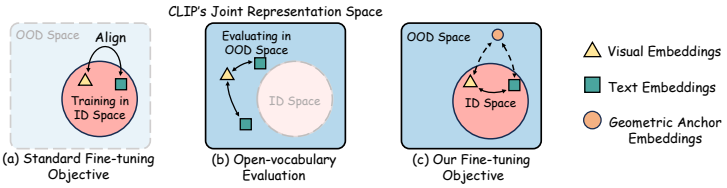


Figure 1: (a) Standard fine-tuning aims to align visual and text embeddings within the known training distribution (ID Space), but neglects the deviation in the out-of-distribution space. While (b) the open-vocabulary model is evaluated in the OOD space, (c) our fine-tuning objective can effectively maintain the relative structure across the entire embedding space and improve the generalization.

Intuitively, without explicit constraints on the out-of-distribution (i.e. OOD) space, visual and text representations may drift in different directions, thereby disrupting their cross-modal alignment. Under the open-vocabulary setting, such alignment disruption can directly undermine the model's generalization to videos from unseen categories. In practice, since the fine-tuning dataset is limited in both scale and diversity, this inconsistency between fine-tuning and evaluation objectives is prevalent in existing adaptation paradigms, as illustrated in Figure 1. We revisit this inconsistency to better understand what must be preserved for open-vocabulary generalization. Our analysis suggests that OOD generalization still largely relies on CLIP's original semantic space, whereas standard fine-tuning often induces representation drift and cross-modal alignment shift beyond the training distribution, leading to degraded generalization.

Building upon these findings, we present *TACO*, a simple yet effective framework for task-consistent open-vocabulary video adaptation, and formulate a more explicit adaptation principle. Our key insight is that the adaptation process should properly regularize the entire representation space, thereby aligning more closely with the evaluation objective and curbing harmful alignment shift in OOD space. To this end, we propose *Relative Structure Distillation*, which ensures the consistent structure of the entire representation space throughout fine-tuning in a data-independent manner. Unlike existing distillation objectives that focus on matching within training distribution, our objective connects the ID space with a constructed OOD space, thereby effectively suppressing drift in OOD regions. We further show that simple random geometric anchors sampled on the CLIP hypersphere are sufficient to provide effective global structural references. Second, we propose a lightweight *Specialization Projection* that decouples the representation space from the optimization space, so that task-specific adaptation can occur in a disposable branch rather than directly overspecializing the representations used at test time. These two designs directly follow our analysis: RSD protects the global alignment structure beyond the training categories, while the specialization projection reduces the tendency of the cross-entropy objective to overfit the shared representation space. Overall, *TACO* is concise, scalable, and easily integrated into different adaptation pipelines, delivering significant improvements. Extensive experiments show that TACO consistently improves open-vocabulary generalization and establishes new state-of-the-art performance across multiple datasets and backbone scales.

Our main contributions can be summarized as follows:

- We introduce *TACO*, a simple yet effective framework for open-vocabulary video adaptation. *TACO* aims to mitigate the potential negative effects induced by the representation deviation beyond the training distribution, a factor not directly addressed in prior work.
- We revisit the objective inconsistency in existing fine-tuning paradigms, analyze its impact on preserving pretrained generalization, and formulate a more explicit adaptation principle (§ 2).
- We present Relative Structure Distillation, which effectively preserves generalization by keeping the consistent structure of OOD space without explicit OOD supervision. We further propose a generic representation-space decoupling scheme to mitigate overspecialization (§ 3).
- *TACO* sets state-of-the-art performance on cross-dataset and base-to-novel settings across multiple benchmarks. Comprehensive ablations further demonstrate the effectiveness of our method (§ 4).

2 Analysis: Revisiting Open-Vocabulary Adaptation for Video

2.1 Preliminary: Adapting CLIP for Video Recognition

To transfer the powerful generalization of the CLIP model [37] to the video domain, recent studies have fine-tuned its joint embedding space using video-text data and obtained promising results [33,

38, 55]. We briefly describe the standard fine-tuning paradigm in this context. Consider a CLIP-based video learner composed of a visual encoder $f_{\theta_v}(\cdot)$ and a text encoder $f_{\theta_t}(\cdot)$, where θ_v and θ_t are the parameters of each encoder. Given a video $V \in \mathbb{R}^{T\times H\times W\times 3}$ with T frames and the corresponding text description C embedded in a set of pre-defined templates (e.g., *"A video of []"*), we extract the visual embedding $\boldsymbol{v} \in \mathbb{R}^D$, text embedding $\boldsymbol{c} \in \mathbb{R}^D$, and compute their similarity $\mathrm{sim}(\boldsymbol{v}, \boldsymbol{c})$ as:

$$\boldsymbol{v} = f_{\theta_v}(V), \;\; \boldsymbol{c} = f_{\theta_t}(C), \;\; \mathrm{sim}(\boldsymbol{v}, \boldsymbol{c}) = \frac{\langle \boldsymbol{v}, \boldsymbol{c} \rangle}{\|\boldsymbol{v}\| \, \|\boldsymbol{c}\|}, \tag{1}$$

where D is the dimension of the joint embedding space. During fine-tuning, video-specific knowledge is injected by encouraging the video embedding to align with the matched category text embedding while separating it from other categories. Formally, the training objective can be written as:

$$\mathcal{L}_{CE} = \mathbb{E}_{(V,C)\sim\mathcal{D}}[\, CE(\, \mathrm{sim}(\boldsymbol{v}, \boldsymbol{W}),\, onehot(C)\,)\,], \tag{2}$$

where $\mathcal{D}$ is the fine-tuning dataset, $\boldsymbol{W} \in \mathbb{R}^{N\times D}$ denotes the text embeddings of N training categories, $CE(\cdot,\cdot)$ is the cross-entropy loss with softmax operation, and $onehot()$ denotes the one-hot encoding.

2.2 Inconsistency Between Open-vocabulary Adaptation and Evaluation

Following the above fine-tuning paradigm, great efforts have been made to adapt CLIP for open-vocabulary video recognition [15, 42, 49, 55]. These methods typically conduct fine-tuning on the Kinetics-400 [20] dataset, and the derived model is expected to generalize well on test data with unseen categories $C_{test} \in \mathcal{S}_{test}$, where $|\mathcal{S}_{ft} \cap \mathcal{S}_{test}| < |\mathcal{S}_{ft} \cup \mathcal{S}_{test}|$. $\mathcal{S}_{ft}$ and $\mathcal{S}_{test}$ denote the fine-tuning and test vocabularies. However, the standard fine-tuning objective is optimized restrictively within the in-distribution (ID) training space, neglecting the basic setting of evaluating open-vocabulary tasks in the OOD space. In this section, we analyze the potential negative impact of such inconsistency in preserving generalization, supported by empirical evidence.

A closer look at preserving generalization. To better understand the essence of preserved generalization, we investigate the impact of each fine-tuned encoder on generalization by replacing it with the corresponding CLIP's original encoders during evaluation. We show the harmonic mean of zero-shot performances in Figure 2 (a). Interestingly, replacing the fine-tuned text encoder with CLIP's text encoder has almost no impact on the results (66.3% vs. 66.2%), which suggests that the OOD generalization of the adapted model is grounded in CLIP's original semantic space. However, fine-tuning inevitably leads to deviations in the OOD representation space (evidenced in Figure 3), as the representations of the test data are not orthogonal to the ID subspace spanned by the training data [24]. This deviation may further distort the representation alignment between visual and text modalities in OOD space and result in diminished generalization capability. Based on these findings, we hypothesize that *maintaining consistent representation alignments in OOD space is crucial for preserving generalization.*

	FT Text	CLIP Text
FT Visual	**66.3**	**66.2**
CLIP Visual	61.4	60.9

(a) Standard FT model

	FT Text	CLIP Text
FT Visual	68.4	67.3
CLIP Visual	62.8	60.9

(b) Our FT model

Figure 2: Replacing the encoders of the standard fine-tuning model and our model with the original CLIP encoders.

Characterizing representation deviation in OOD Space. To support the above hypothesis, we investigate how representations deviate in OOD space during fine-tuning. Specifically, Figure 3 visualizes the similarity distributions between visual embeddings from the CLIP model and various fine-tuned models, with the averaged overall text embedding similarity shown in the legend. From the figure, a pronounced deviation can be observed in visual embeddings for standard fine-tuning paradigms. Interestingly, enabling the text encoder to be updated can effectively suppress this deviation. This phenomenon reveals that the text encoder serves as a critical regularizer to *relieve the overfitting in visual representations, rather*

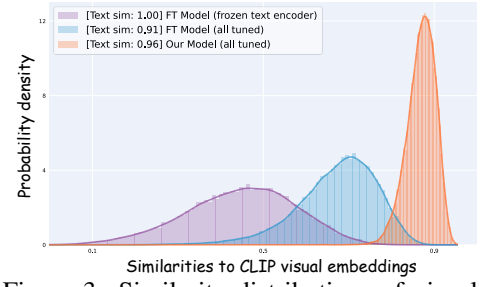


Figure 3: Similarity distributions of visual embeddings between the CLIP model and various fine-tuned models in OOD space (UCF, HMDB, and K600).

than learning new textual knowledge. However, there is still a noticeable deviation for both fine-tuned visual and text embeddings that degrades the OOD generalization. We believe such deviation should be further curbed during fine-tuning, which was overlooked in previous studies.

Quantifying alignment shift in adaptation. Beyond the deviation of individual representations in OOD space, a more critical issue is whether such deviation further disrupts the original cross-modal alignment. We therefore quantify the resulting alignment shift and examine its relationship with the generalization performance HM_{ood}. Specifically, we use the visual-text similarity matrix of the original CLIP in OOD space as the reference alignment, and measure the shift of an adapted model by the KL divergence D_{KL} between their similarity matrices. A larger D_{KL} indicates a greater departure from the original alignment. As shown in Figure 4, a moderate degree of alignment update benefits generalization by bridging the image–video domain gap, whereas *excessive shifts correlate with degraded generalization due to cross-modal disruption*. This finding offers a new lens for interpreting the efficacy of existing methods by highlighting their role in protecting OOD alignment. However, existing methods generally suffer from over-shifting, as they overlook the importance of constraining the variation in OOD space.

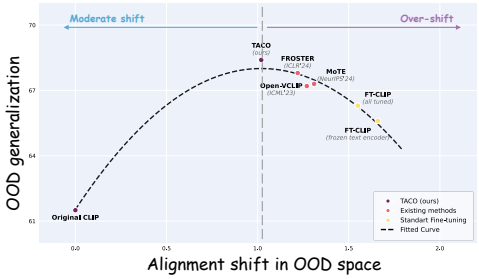


Figure 4: Alignment shift D_{KL} and generalization performance of various adapted models in OOD space (UCF, HMDB, and K600)

3 Methodology

3.1 Motivation and Intuition

Building upon the above analysis, we argue that maintaining consistent representation alignments in the OOD space is critical for generalizable VLM adaptation. Although existing methods can shape a well-structured ID space, their optimization objectives remain confined to the training distribution, offering limited constraint over representation variation beyond it and thus leading to suboptimal adaptation. A straightforward remedy would be to explicitly model the OOD space using real paired OOD vision-language data. In practice, however, this is extremely challenging, as it requires collecting large amounts of real OOD data and incurs substantial computational overhead. This challenge inspires us to seek an indirect yet effective alternative: rather than explicitly modeling OOD semantics, we regularize the alignment of the entire representation space by preserving the **relative structure** between the ID and constructed OOD spaces. In this way, we can bypass the expensive collection and computational costs required by direct OOD modeling. This intuition naturally leads to our *Relative Structure Distillation*, a novel regularizer that enforces geometric structural consistency with the pre-trained teacher beyond the training distribution, mitigating harmful alignment shifts without explicit OOD supervision, as presented in Figure 5. From a geometric perspective, RSD can be viewed as random-projection-based relational regularization: the anchors probe structural changes without requiring advance knowledge of which directions correspond to unseen semantic concepts.

3.2 Relative Structure Distillation

In Section 2.2, we reveal that the preserved generalization of an adapted model is rooted in CLIP's original semantic space. Building on this insight, the core challenge during fine-tuning is to maintain consistency with the pretrained reference alignment in the OOD space. Naturally, this challenge can be cast as a knowledge distillation problem. However, existing distillation strategies—ranging from standard Kullback-Leibler (KL) divergence [1] and L2 matching [15] to relation-based variants [53]—are fundamentally limited in this setting. Their distillation objectives typically focus on matching within the training distribution since the distillation data is identical to the training data, leaving them ineffective in suppressing deviations beyond it. In contrast, our proposed *Relative Structure Distillation* introduces principled solutions to the above challenge via refinements in loss formulation, distillation data, and optimization space.

Formulation of Relative Structure Distillation. Our design inspiration is to extend the regularization scope to the entire representation space by establishing connections between the training

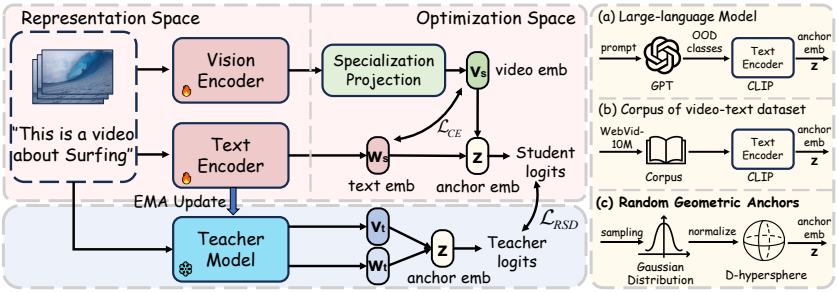


Figure 5: An overview of the TACO framework (left) and the source of geometric anchors (right).

distribution and *a constructed OOD representation space*. Formally, given a video and a set of categories, we obtain the normalized teacher visual embedding $\boldsymbol{v_t} \in \mathbb{R}^{1\times D}$ and teacher text embeddings $\boldsymbol{W_t} \in \mathbb{R}^{N\times D}$, along with the student embeddings $\boldsymbol{v_s}, \boldsymbol{W_s}$ of the same dimensions. The teacher model shares the same structure as the student model. Then we sample M normalized anchor representations $\boldsymbol{Z} \in \mathbb{R}^{M\times D}$ from the constructed OOD space. We take $\boldsymbol{Z}$ as geometric anchors to enforce the consistency of its relative relationships with the visual and text embeddings in ID space:

$$\mathcal{L}_{RSD} = D_{KL}\Big(\sigma\big(\boxed{\boldsymbol{v}_t\boldsymbol{Z}^\top}\ \boxed{\boldsymbol{Z}\boldsymbol{W}_t^\top}/\tau\big) \,\big\|\, \sigma\big(\boxed{\boldsymbol{v}_s\boldsymbol{Z}^\top}\ \boxed{\boldsymbol{Z}\boldsymbol{W}_s^\top}/\tau\big)\Big), \tag{3}$$

where σ denotes the softmax operation, τ is a temperature parameter. Here, our goal is to protect the *geometric structure* of the entire representation space rather than modeling the truly out-of-distribution semantics. The geometric anchors $\boldsymbol{Z}$ capture the relations with the visual embedding and text embedding through the terms $\boldsymbol{v}\boldsymbol{Z}^\top$ and $\boldsymbol{Z}\boldsymbol{W}^\top$. During fine-tuning, if the specialized CE loss overemphasizes the agreement between $\boldsymbol{v}$ and $\boldsymbol{W}$ in the ID space, their relations to the constructed OOD space may drift away. The anchor-mediated relations $\boldsymbol{v}\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{W}^\top$ can capture such relative structural changes and steer the student back toward a teacher-consistent geometry.

Appendix D provides further theoretical support for the random-anchor formulation of RSD. Standard KL matching can leave cross-modal relation discrepancies invisible to ID logits. For any nonzero class-pair relation discrepancy in this null space, isotropically sampled anchors produce a positive expected squared response, providing additional geometric measurements beyond direct logit matching.

Construction of the OOD space. The geometric anchors $\boldsymbol{Z}$ are designed to serve as geometric reference points in the representation space, providing a geometric scaffold for preserving the structure of the entire space. In this regard, the anchor embeddings sampled from the constructed OOD space do not need to be semantically associated with the training samples. However, for such anchors to serve as meaningful geometric references, the constructed OOD space should (**i**) reside in the same joint representation space as the model (*compatibility*) and (**ii**) ideally span the entire semantic space (*coverage*). One possible way is to sample from a large authentic corpus. For example, leveraging the textual descriptions in existing large-scale video-text datasets (e.g., WebVid-10M [2]) or prompting large-language models [16] to generate potential video categories. But the sampled texts may not uniformly span the semantic space and require additional computations for embedding.

To address this, we propose *Random Geometric Anchors*, which directly sample anchor embeddings from the unit D-hypersphere of the CLIP model, where D denotes the embedding dimension. Specifically, we draw samples from the standard Gaussian distribution $\mathcal{N}(0, I)$ and project them onto the unit hypersphere by normalization. Owing to the rotational invariance of the Gaussian distribution, the resulting anchors are uniformly distributed on the hypersphere. As previous work has shown that CLIP's representation space is confined within an extremely narrow conical region on the D-hypersphere [26], the hypersphere can be regarded as a generalized OOD space, and the probability of overlap between the sampled anchors and the ID space is practically negligible during fine-tuning. Therefore, although these anchors do not carry authentic semantic meaning, they naturally satisfy the desired properties of compatibility and coverage, and can effectively characterize the relative structure between ID and OOD spaces. Moreover, these anchors should not be viewed as mere noise, but rather as geometric reference points that support structure preservation. This is analogous to preserving a map's layout with reference landmarks: even if the landmarks themselves are not meaningful destinations, they still help maintain the relative structure of the entire space.

The utility of these anchors does not require high cosine similarity to individual semantic directions. For a uniformly sampled unit anchor z, $\mathbb{E}[zz^\top] = I/D$, giving no preferred direction in expectation.

Resampling introduces new geometric probes across iterations, with their sensitivity to relation discrepancies analyzed in Appendix D.

Design of the teacher model. While the frozen CLIP provides a highly stable alignment reference as the teacher model, we observe that its distillation supervision may hinder the model's ability in learning new knowledge. In our work, the teacher model is designed as an exponential moving average (EMA) model with its parameters slowly updated at each iteration as $\tilde{\theta} = \alpha\tilde{\theta} + (1-\alpha)\theta$, where α is a momentum coefficient, $\tilde{\theta}$ and θ are the parameters of the teacher and student models. This mechanism ensures the teacher model remains consistent while evolving with new knowledge. The overall learning objective can be defined as follows, with $\lambda = 0.4$ by default:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda\,\mathcal{L}_{RSD}, \tag{4}$$

3.3 Decoupling of the Representation Space

In standard fine-tuning, the representation space is typically identical to the optimization space, as the output visual embeddings are directly used to compute the training objective. However, the narrow supervision from a limited set of text categories tends to pull CLIP's originally open and transferable visual representations toward a skewed subspace specialized for the training distribution. As a result, the very space that should generalize to unseen categories becomes directly exposed to task-specific bias. While Relative Structure Distillation explicitly regularizes the geometry of the representation space, it does not prevent the video-specific CE loss from directly specializing that space. Therefore, we further introduce a mechanism to protect the representation space from such direct specialization pressure, thereby mitigating harmful representation drift.

To this end, we propose to decouple the representation space from the optimization space by simply inserting a lightweight *Specialization Projection* between them. Specifically, given a video input V, we first obtain its visual representation $f_{\theta_v}(V)$ and then transform it into the optimization space with a projection network $h(\cdot)$ as $\hat{\boldsymbol{v}} = h(f_{\theta_v}(V)) + f_{\theta_v}(V)$. Here $\hat{\boldsymbol{v}} \in \mathbb{R}^D$ is the projected visual embedding used for optimization, while $f_{\theta_v}(V)$ is retained as the underlying representation. To delicately balance the effective decoupling effect while keeping the alignment between text and pre-projected visual embeddings, we adopt a residual linear projection with zero-initialized parameters [13] to avoid abrupt shifts at the start of the fine-tuning. We further remove non-linear activations, as they tend to cause harmful alignment changes when the text encoder is jointly optimized, and use a larger learning rate for the projection head than for the visual encoder to strengthen the decoupling effect.

During evaluation, we discard the projection head for more generalizable representations. Our goal here is not merely generic overfitting reduction, but to prevent the representation space from being directly specialized by the CE objective. Unlike conventional regularizations such as weight decay or dropout, our decoupling strategy is an architectural bias that allows specialization to occur in a disposable optimization branch, rather than in the representation space that must generalize to unseen categories. Thus, it is particularly well-matched to our goal of protecting the space structure.

4 Experiments

4.1 Experimental Setup

Evaluation Protocols. We thoroughly evaluate our method with two common protocols: *Cross-dataset* and *Base-to-novel* evaluation [15, 38]. ***(i)*** *Cross-dataset:* In this setup, the model is trained on Kinetics-400 [20] and then evaluated on other datasets with out-of-distribution vocabulary, including UCF-101 [39], HMDB-51 [23], and Kinetics-600 [3]. For UCF and HMDB, we evaluate on both the full dataset and three validation splits. For K600, we adopt the three splits provided by [4]. Each split contains 160 categories sampled from 220 unseen categories. ***(ii)*** *Base-to-novel:* In this protocol, a dataset is divided into two disjoint category sets: base classes and novel classes. The model is tuned on base classes and evaluated on both base and novel classes. Evaluation datasets including K400, UCF, HMDB, and Something-Something v2 [10]. During inference, we sample 3 temporal clips with a center crop (i.e. 3×1 views) per video, with each view consisting of 8 frames.

Implementation Details. We use the CLIP [37] pre-trained ViT-B/16 and ViT-L/14 models in our experiments. During fine-tuning, we sparsely sample 8 frames as the video input. The pre-processing

Table 1: Ablations on the components and key details. We report the cross-dataset performance on UCF, HMDB, and K600 split1, using the ViT-B/16 network. UCF$_f$ and K600$_f$ denote the results of freezing the text encoder in fine-tuning. Default settings are colored in blue.

(a) Effects of each components.

Spec.	Proj.	$\mathcal{L}_{\text{RSD}}$	UCF	HMDB	K600
			82.8	52.0	73.4
✓			83.7	53.2	74.8
	✓		84.5	52.9	75.2
✓	✓		**84.9**	**54.2**	**75.7**

(b) Source of geometric anchors.

Methods	UCF	HMDB	K600
LLM	83.7	53.2	74.6
WebVid-10M	84.2	53.8	75.5
Random $ZZ^\top$	75.2	46.8	69.2
$\mathcal{N}(0, I)$	**84.9**	**54.2**	**75.7**

(c) Various distillation objectives.

Type	UCF	HMDB	K600
L2 Loss	83.7	53.1	75.1
KL divergence	83.8	53.6	75.0
Relation NCE	83.9	53.0	74.7
$\mathcal{L}_{\text{RSD}}$	**84.9**	**54.2**	**75.7**

(d) Number of the random geometric anchors M in distillation.

Number	UCF	HMDB	K600
2	83.9	53.0	74.7
20	84.6	53.9	75.4
200	**84.9**	**54.2**	**75.7**
2000	84.0	53.8	75.6

(e) Various implementations of the Specialization Projection.

Type	UCF	K600	UCF$_f$	K600$_f$
None	82.8	73.4	81.5	69.7
MLP	82.2	73.3	**83.1**	71.9
Transformer	81.2	72.4	82.7	71.5
Two linear	**83.7**	**74.8**	82.8	**72.1**

(f) Effects of applying rephrased text and weight ensemble.

Methods	UCF	HMDB	K600
None	84.9	54.2	75.7
+Rephrased text	**86.0**	54.3	77.4
+Weight ensemble	85.9	**54.6**	**78.1**

Table 2: Structural preservation on unseen categories relative to the original CLIP space. Dir. cos.: text-direction cosine; kNN@5: text kNN overlap@5; ρ: cross-modal Spearman correlation.

Method	UCF			HMDB			K600		
	Dir. cos. ↑	kNN@5 ↑	ρ ↑	Dir. cos. ↑	kNN@5 ↑	ρ ↑	Dir. cos. ↑	kNN@5 ↑	ρ ↑
Standard fine-tuning	0.87	0.65	0.75	0.88	0.70	0.76	0.90	0.65	0.79
TACO (ours)	**0.95**	**0.72**	**0.84**	**0.94**	**0.77**	**0.83**	**0.95**	**0.75**	**0.84**

includes random cropping and resizing to the size of 224×224, along with random horizontal flips and grayscale. We adopt AdamW [29] optimizer with a weight decay of 0.2. The initial learning rate is set to 3.75×10^{-6} with a total batch size of 192, following a half-period learning rate decay. For specialization projection parameters, the initial learning rate is increased to 5×10^{-5}. Furthermore, we set the anchor embedding number M to 200 and the momentum coefficient α to 0.9998. The geometric anchors are resampled at each training iteration. *Please see supplementary for more details.*

4.2 Ablation Studies

Component-wise analysis of *TACO*. To analyze the effect of proposed components, we perform in-depth ablations with the ViT-B/16 network in Table 1a. Our baseline is the ViT-B/16 CLIP model adapted under the standard fine-tuning paradigm. The results shows that both Specialization Projection and $\mathcal{L}_{\text{RSD}}$ are effective in mitigating harmful alignment shift from two complementary mechanisms: decoupled optimization pathway and structural regularization, respectively. Their combination yields a promising synergistic effect, demonstrating the effectiveness of our method.

Construction of the geometric anchors. We compare different anchor construction strategies in Table 1b, including semantic anchors derived from LLM-generated video categories [16] and WebVid text descriptions [2], with embeddings reduced by K-means clustering [31]. The results show that our Gaussian sampling strategy performs better while requiring no additional text collection or embedding computation. This suggests that the gain of $\mathcal{L}_{\text{RSD}}$ does not primarily come from semantic information carried by the anchors, but from providing uniformly distributed geometric references in the shared embedding space. Compared with semantically meaningful but distributionally biased anchors, random geometric anchors offer more balanced global coverage. We further replace the anchor term in Eq. 3 with a randomly initialized $D \times D$ matrix, which causes training collapse. This shows that the gain does not come from arbitrary randomness. Since a random matrix does not lie in the CLIP hypersphere, it introduces unconstrained transformations with no structural meaning. Therefore, effective anchors must be valid points in the same representation space as the model.

Effects of regularizing the OOD space. In Table 1c, we compare L2, KL divergence, and a relation-based NCE objective [53] built from matched relation pairs and queue-based negatives. The results show that preserving the structure of the entire representation space is crucial for generalization. By regularizing the geometry induced by the constructed OOD anchors, our method better suppresses representation deviation beyond the training distribution and achieves superior performance.

Table 3: Performance comparison with state-of-the-art methods under the base-to-novel setting. "HM" indicates the harmonic mean of the accuracy on base and novel sets.

Method	Venue	K400			HMDB			UCF			SSv2		
		BASE	Novel	HM	BASE	Novel	HM	BASE	Novel	HM	BASE	Novel	HM
CLIP [37]	ICML'21	62.3	53.4	57.5	53.3	46.8	49.8	78.5	63.6	70.3	4.9	5.3	5.1
ActionCLIP [40]	arXiv'21	61.0	46.2	52.6	69.1	37.3	48.5	90.1	58.1	70.7	13.3	10.1	11.5
X-CLIP [33]	ECCV'22	74.1	56.4	64.0	69.4	45.5	55.0	89.9	58.9	71.2	8.5	6.6	7.4
VPT [19]	ECCV'22	69.7	37.6	48.8	46.2	16.0	23.8	90.5	40.4	55.8	8.3	5.3	6.4
AIM [48]	ICLR'23	74.6	62.5	68.0	64.0	51.6	57.1	89.8	76.4	82.6	8.5	7.9	8.2
ST-Adapter [34]	NeurIPS'22	73.6	62.0	67.3	65.3	48.9	55.9	85.5	76.8	80.9	9.3	8.4	8.8
ViFi-CLIP [38]	CVPR'23	76.4	61.1	67.9	73.8	53.3	61.9	92.9	67.7	78.3	16.2	12.1	13.9
Open-VCLIP [42]	ICML'23	76.5	62.6	68.9	70.3	50.4	58.7	94.8	77.5	85.3	16.0	11.0	13.0
FROSTER [15]	ICLR'24	77.8	**64.3**	**70.4**	74.1	58.0	65.1	95.3	80.0	87.0	**18.3**	12.2	**14.6**
Open-MeDe [49]	ICCV'25	77.2	63.8	69.9	73.6	56.4	63.9	94.9	78.5	85.9	17.1	12.3	14.3
TACO (ours)		**78.2**	63.8	70.3	**74.4**	**62.3**	**67.8**	**95.8**	**82.2**	**88.5**	17.7	**12.4**	**14.6**

Varying numbers M of random geometric anchors. Table 1d shows that increasing M improves results by providing better coverage of the constructed OOD space, and $M = 200$ is sufficient to capture the relative structure relationships. Further increasing M slightly hurts performance, since the distillation term becomes overly restrictive and hinders task-specific adaptation.

Various implementations of the specialization projection. Table 1e compares different specialization projection designs under two fine-tuning settings, depending on whether the text encoder is updated. When the text encoder is frozen, all three designs improve performance; however, when it is tunable, MLP and Transformer projections degrade generalization. This suggests that the optimization branch should remain close to the original representation space: more expressive projections may over-modify the optimization embedding, especially when the text encoder is also updated, making cross-modal alignment harder to preserve. In contrast, two simple linear layers effectively decouple the representation and optimization spaces, yielding consistent gains in both settings.

Effects of the rephrased text and weight ensemble. To further enhance generalization, we incorporate the rephrased text descriptions from FROSTER [15] and the weight ensemble technique from Open-VCLIP [42]. As shown in Table 1f, both are orthogonal to TACO and can provide additional gains. Notably, TACO already achieves clear improvements without relying on these auxiliary techniques, while rephrased prompts or weight ensembling may further help depending on the training dynamics. In particular, we find that weight ensembling is most effective when the model is over-trained, but may be less beneficial otherwise.

Structural preservation on unseen categories. To complement the accuracy comparison, we directly examine structural preservation on real unseen categories using the original CLIP space as a reference. As shown in Table 2, across UCF, HMDB, and K600, TACO consistently improves text-direction cosine, text kNN overlap@5, and cross-modal Spearman correlation over standard fine-tuning. These results provide complementary evidence that TACO better preserves unseen-category directions, local neighborhoods, and cross-modal rankings. Metric definitions are provided in Appendix E. *Please see supplementary for more ablations.*

4.3 Main Results

Base-to-novel video recognition. Table 3 reports the results under the base-to-novel setting, which reflects the model's joint ability to fit video-specific biases while being adapted to unknown categories. TACO achieves the best or competitive harmonic-mean performance across all four benchmarks, with especially clear gains on HMDB and UCF. Notably, the improvements are mainly driven by better recognition of novel classes, suggesting that our adaptation does not simply bias the model toward the training categories. This observation is consistent with our method design: by preserving the broader geometric structure of the pretrained space, TACO enables video adaptation while better maintaining the alignment needed for unseen categories. The gains on SSv2 are more modest, likely because SSv2 requires stronger temporal reasoning, whereas our method mainly targets task-consistent open-vocabulary adaptation. Still, TACO remains competitive without introducing additional temporal modules, indicating that it is complementary to dedicated temporal modeling techniques.

Cross-dataset video recognition. Table 4 presents comparisons with the state-of-the-art methods under the cross-dataset setting, which assesses the model's generalization towards out-of-distribution

Table 4: Zero-shot classification performance compared with the state-of-the-art methods under the cross-dataset setting, evaluated on the *validation splits* of UCF-101, HMDB-51, and Kinetics-600.

Method	Venue	Encoder	Frames	UCF-101	HMDB-51	Kinetics-600
ActionCLIP [40]	arXiv'21	ViT-B/16	32	58.3±3.4	40.8±5.4	67.7±1.1
X-CLIP [33]	ECCV'22	ViT-B/16	32	72.0±2.3	44.6±5.2	65.2±0.4
ST-Adapter [34]	NeurIPS'22	ViT-B/16	8	76.9±0.8	51.5±0.6	60.2±1.8
Vita-CLIP [41]	CVPR'23	ViT-B/16	8/32	75.0±0.6	48.6±0.6	67.4±0.5
ViFi-CLIP [38]	CVPR'23	ViT-B/16	32	76.8±0.7	51.3±0.6	71.2±1.0
Open-VCLIP [42]	ICML'23	ViT-B/16	8	83.4±1.2	53.9±1.2	73.0±0.8
MAXI [27]	ICCV'23	ViT-B/16	16/32	78.2±0.8	52.3±0.7	71.5±0.8
FROSTER [15]	ICLR'24	ViT-B/16	8	84.8±1.1	54.8±1.3	74.8±0.9
OST [5]	CVPR'24	ViT-B/16	8	77.9±1.3	54.9±1.1	73.9±0.8
MoTE [55]	NeurIPS'24	ViT-B/16	8	83.4±0.7	55.8±0.9	70.2±0.6
Open-MeDe [49]	ICCV'25	ViT-B/16	8	83.7±1.3	54.6±1.1	73.7±0.9
TACO (ours)		ViT-B/16	8	**85.6**±1.2	**60.0**±0.5	**77.0**±0.9
X-Florence [33]	ECCV'22	Florence	32	73.2±4.2	48.4±4.9	68.8±0.9
Text4Vis [44]	AAAI'23	ViT-L/14	8	82.6±0.7	52.4±0.4	72.1±0.9
OTI [56]	ACMMM'23	ViT-L/14	8	88.1±1.0	59.3±1.7	70.6±0.5
Open-VCLIP [42]	ICML'23	ViT-L/14	8	87.6±1.2	59.0±0.6	81.1±0.8
DiST [36]	ICCV'23	ViT-L/14	32	74.9±0.8	57.5±1.6	75.0±0.7
MoTE [55]	NeurIPS'24	ViT-L/14	8	88.7±0.6	61.4±1.3	78.4±0.9
TACO (ours)		ViT-L/14	8	**91.4**±0.7	**64.2**±0.8	**83.9**±0.7

Table 5: Zero-shot performance of UCF and HMDB on the *full* dataset. * denotes evaluation with the full validation set on HMDB.

Method	Encoder	UCF	HMDB
CLIP [37]	ViT-B/16	74.9	46.7
AIM [48]	ViT-B/16	79.0	49.5
ST-Adapter [34]	ViT-B/16	77.9	50.3
Open-VCLIP* [42]	ViT-B/16	83.5	53.2
FROSTER* [15]	ViT-B/16	85.0	54.5
TACO (ours)	ViT-B/16	**85.9**	**54.6**
Text4Vis [44]	ViT-L/14	79.6	49.8
BIKE [44]	ViT-L/14	80.8	52.8
OTI [56]	ViT-L/14	88.3	55.8
MoTE [55]	ViT-L/14	89.4	56.3
TACO (ours)	ViT-L/14	**91.6**	**59.9**

Table 6: Integrating our TACO with various adaptation methods can significantly improve the zero-shot generalization.

Type	Method	UCF	HMDB	K600
Adapter-based	ST-Adapter [34]	77.9	50.3	60.2
	+ *TACO(ours)*	**80.5**	**53.1**	**73.9**
	Δ	**+2.6**	**+2.8**	**+13.7**
Prompt-based	Vita-CLIP [41]	78.6	50.5	67.4
	+ *TACO(ours)*	**80.3**	**52.6**	**73.2**
	Δ	**+1.7**	**+2.1**	**+5.8**
LoRA-based	CLIP + LoRA [13]	80.5	50.4	71.2
	+ *TACO(ours)*	**83.1**	**54.4**	**74.7**
	Δ	**+2.6**	**+4.0**	**+3.5**
Fully-tuned	Fine-tuned CLIP	82.8	52.0	73.4
	+ *TACO(ours)*	**84.9**	**54.2**	**75.7**
	Δ	**+2.1**	**+2.2**	**+2.3**

categories. Under this protocol, TACO consistently outperforms prior methods across all datasets and both backbone scales, and the same advantage is retained on the full-dataset evaluation in Table 5. These gains are consistent with Figure 3 and Figure 4: TACO reduces representation drift in OOD regions and maintains a more moderate visual-text alignment shift during adaptation, both of which are important for open-vocabulary generalization. This matches our design, where $\mathcal{L}_{\text{RSD}}$ regularizes relations beyond the training categories and the Specialization Projection reduces over-specialization of the representation space. As a result, TACO better preserves the transferable structure inherited from CLIP, especially under larger distribution shifts such as HMDB and Kinetics-600.

Applicability to various adaptation methods. Table 6 further shows that TACO can be integrated with adapter-based, prompt-based, LoRA-based, and fully fine-tuned adaptation methods, consistently improving their zero-shot generalization. This suggests that our framework is largely orthogonal to the underlying adaptation mechanism, since it acts by preserving representation geometry and decoupling it from the optimization space. The strongest absolute performance is obtained with full fine-tuning, suggesting that broader parameter updates provide greater capacity for video adaptation, while our method helps prevent this flexibility from causing excessive drift in the pretrained space.

5 Visualizations

To obtain a more intuitive understanding of the representation distribution, we present t-SNE [30] visualizations of the visual representations from CLIP, Fine-tuned CLIP, and our model in Figure 6. The representations are extracted from the UCF-101 dataset. As shown in the figure, the representation distribution of the CLIP model is relatively messy due to a lack of video-specific knowledge. Compared with the standard fine-tuned model, our TACO exhibits a more compact intra-class distribution and more pronounced inter-class distinctions. This phenomenon demonstrates that our approach can introduce new video knowledge while preserving the relative structure among unknown categories,

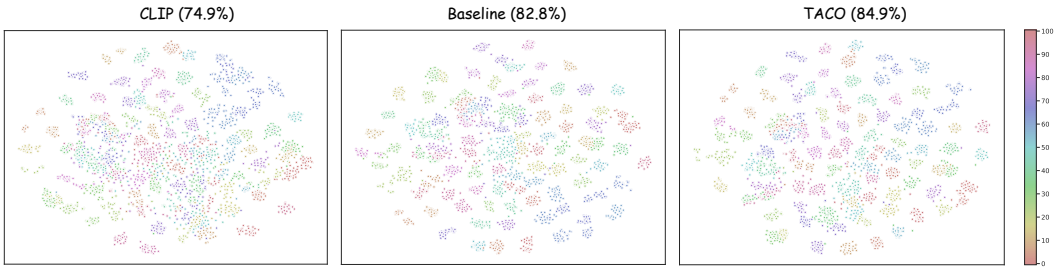


Figure 6: t-SNE [30] visualization of UCF-101 visual representations from CLIP, standard fine-tuning, and TACO (left to right).

thereby achieving greater performance improvements, which can be attributed to the introduction of Relative Structure Distillation.

6 Related Work

Adapting VLMs to video recognition. Adapting VLMs to video recognition has proven effective, but remains challenging in balancing video-specific adaptation with the pretrained generalization of VLMs [38]. To better capture temporal dynamics, existing methods introduce additional parameterized modules, such as adapters [34, 48], prompts [18, 19, 41], and Transformer layers [44, 45, 55]. For example, X-CLIP [33] proposes an attention-based multi-frame integration module for cross-frame information exchange. Another line of work focuses on preserving generalization during adaptation: Open-VCLIP [42] interpolates model weights along the optimization trajectory, FROSTER [15] alleviates the overfitting by ensuring the learned features do not diverge too far from the frozen CLIP, and Open-MeDe [49] uses meta-optimization to mitigate the static bias of the pretrained model. Despite their strong open-vocabulary performance, these methods still optimize mainly within the training distribution, leaving the mismatch between adaptation and open-vocabulary evaluation insufficiently addressed. Our method revisits this issue from the perspective of representation consistency beyond the training distribution and achieves stronger generalization.

Knowledge distillation for VLMs. Knowledge distillation has been widely used to adapt pretrained models and improve generalization [25, 35]. Its core idea is to transfer the teacher's knowledge by matching logits, features, or intermediate representations, either to improve the student's generalization [7, 32, 46] or constrain fine-tuning with pretrained supervision [1, 15]. For example, CLIPPING [35] distills layer-wise knowledge from a larger teacher into an efficient student model, while FROSTER [15] reduces overfitting through residual feature distillation. Relation-based distillation methods such as CRCD [53] are also related, but they are designed for closed-set classification and transfer local relations among training samples, whereas our method uses random geometric anchors to regularize the global geometry of the entire embedding space for open-vocabulary adaptation. Overall, existing distillation methods mainly operate within the training distribution. In contrast, we emphasize preserving representation consistency beyond the training distribution and propose a distillation objective that maintains the relative structure of the OOD space during adaptation.

7 Conclusion

In this work, we present *TACO*, a concise adaptation framework to mitigate the detrimental effects induced by representation deviations beyond the training distribution. By analyzing the objective inconsistency of existing adaptation paradigms, we formulate a concrete adaptation principle and propose *Relative Structure Distillation* to enforce consistent relative structures across the entire representation space. Furthermore, we introduce a space decoupling strategy with a specialization projection to alleviate the overspecialization. Extensive experiments demonstrate that *TACO* achieves state-of-the-art results across diverse benchmarks under cross-dataset and base-to-novel settings.

Acknowledgements

This paper is supported by the National Natural Science Foundation of China (Grant No. 62333017).

References

[1] Sravanti Addepalli, Ashish Ramayee Asokan, Lakshay Sharma, and R Venkatesh Babu. Leveraging vision-language models for improving domain generalization in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23922–23932, 2024.

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[3] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

[4] Shizhe Chen and Dong Huang. Elaborative rehearsal for zero-shot action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[5] Tongjia Chen, Hongshan Yu, Zhengeng Yang, Zechuan Li, Wei Sun, and Chen Chen. Ost: Refining text knowledge with optimal spatio-temporal descriptor for general video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[6] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4113–4123, 2024.

[7] Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. Enabling multimodal generation on clip via vision-language knowledge distillation. *arXiv preprint arXiv:2203.06386*, 2022.

[8] Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 2017.

[9] Rong Fan, Kaiyan Xiao, Minghao Zhu, Liuyi Wang, Kai Dai, and Zhao Yang. Groundvts: Visual token sampling in multimodal large language models for video temporal grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

[10] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[11] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations (ICLR)*, 2022.

[12] Yuncheng Guo and Xiaodong Gu. Mmrl: Multi-modal representation learning for vision-language models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

[14] Mengxian Hu, Minghao Zhu, Xun Zhou, Qingqing Yan, Shu Li, Chengju Liu, and Qijun Chen. Efficient text-driven motion generation via latent consistency training. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2026.

[15] Xiaohu Huang, Hao Zhou, Kun Yao, and Kai Han. Froster: Frozen clip is a strong teacher for open-vocabulary action recognition. In *International Conference on Learning Representations (ICLR)*, 2024.

[16] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, pages 4904–4916, 2021.

[18] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 105–124, 2022.

[19] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[20] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[21] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[22] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[23] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2556–2563, 2011.

[24] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations (ICLR)*, 2022.

[25] Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[26] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[27] Wei Lin, Leonid Karlinsky, Nina Shvetsova, Horst Possegger, Mateusz Kozinski, Rameswar Panda, Rogerio Feris, Hilde Kuehne, and Horst Bischof. Match, expand and improve: Unsupervised finetuning for zero-shot action recognition with language knowledge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[28] Xiao Lin, Minghao Zhu, Ronghao Dang, Guangliang Zhou, Shaolong Shu, Feng Lin, Chengju Liu, and Qijun Chen. Clipose: Category-level object pose estimation with pre-trained vision-language knowledge. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[30] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008.

[31] James B McQueen. Some methods of classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.*, pages 281–297, 1967.

[32] Marco Mistretta, Alberto Baldrati, Marco Bertini, and Andrew D Bagdanov. Improving zero-shot generalization of learned prompts via unsupervised knowledge distillation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[33] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–18, 2022.

[34] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 26462–26477, 2022.

[35] Renjing Pei, Jianzhuang Liu, Weimian Li, Bin Shao, Songcen Xu, Peng Dai, Juwei Lu, and Youliang Yan. Clipping: Distilling clip-based models with a student base for video-language retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[36] Zhiwu Qing, Shiwei Zhang, Ziyuan Huang, Yingya Zhang, Changxin Gao, Deli Zhao, and Nong Sang. Disentangling spatial and temporal learning for efficient image-to-video transfer learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

[38] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6545–6554, 2023.

[39] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[40] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.

[41] Syed Talal Wasim, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Vita-clip: Video and text adaptive clip via multimodal prompting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23034–23044, 2023.

[42] Zejia Weng, Xitong Yang, Ang Li, Zuxuan Wu, and Yu-Gang Jiang. Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization. In *International Conference on Machine Learning (ICML)*, pages 36978–36989, 2023.

[43] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[44] Wenhao Wu, Zhun Sun, and Wanli Ouyang. Revisiting classifier: Transferring vision-language models for video recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2847–2855, 2023.

[45] Wenhao Wu, Xiaohan Wang, Haipeng Luo, Jingdong Wang, Yi Yang, and Wanli Ouyang. Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6620–6630, 2023.

[46] Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. Clip-kd: An empirical study of clip model distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15952–15962, 2024.

[47] Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. Mma: Multi-modal adapter for vision-language models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[48] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video action recognition. In *International Conference on Learning Representations (ICLR)*, 2022.

[49] Yating Yu, Congqi Cao, Yifan Zhang, and Yanning Zhang. Learning to generalize without bias for open-vocabulary action recognition. *arXiv preprint arXiv:2502.20158*, 2025.

[50] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[51] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 130(9):2337–2348, 2022.

[52] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[53] Jinguo Zhu, Shixiang Tang, Dapeng Chen, Shijie Yu, Yakun Liu, Mingzhe Rong, Aijun Yang, and Xiaohua Wang. Complementary relation contrastive distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[54] Minghao Zhu, Xiao Lin, Ronghao Dang, Chengju Liu, and Qijun Chen. Fine-grained spatiotemporal motion alignment for contrastive video representation learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4725–4736, 2023.

[55] Minghao Zhu, Zhengpu Wang, Mengxian Hu, Ronghao Dang, Xiao Lin, Xun Zhou, Chengju Liu, and Qijun Chen. Mote: Reconciling generalization with specialization for visual-language to video knowledge transfer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[56] Yan Zhu, Junbao Zhuo, Bin Ma, Jiajia Geng, Xiaoming Wei, Xiaolin Wei, and Shuhui Wang. Orthogonal temporal interpolation for zero-shot video recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7491–7501, 2023.

Appendix Overview

This appendix is organized as follows:

- **Section A**: Limitations and Broader Impact.
- **Section B**: More Implementation Details.
- **Section C**: More Ablations.
- **Section D**: Theoretical Analysis of Relative Structure Distillation.
- **Section E**: More Analysis and Discussions.
- **Section F**: Visualizations.
- **Section G**: Textual Prompts Used in Evaluation.
- **Section H**: Dataset Details.

A Limitations and Broader Impact

Limitations While our approach achieves strong performance on open-vocabulary tasks, we observe that its improvements are relatively limited on datasets where temporal information plays a critical role. A promising direction for future research is to integrate our framework with existing temporal modeling techniques to better handle such scenarios. In addition, our method requires setting a relatively small momentum parameter for the teacher model to maintain consistency in the OOD space. However, we observe that the optimal value of this parameter may vary across datasets, which could limit the robustness of our approach. A potential future direction is to develop a more stable strategy for updating the teacher model.

Broader Impact Adapting foundation models to downstream tasks has become a prevailing trend in machine learning [9, 14, 15, 52]. We argue that exploring effective strategies for adapting vision-language models to open-vocabulary tasks is both timely and necessary for real-world applications. This work seeks to offer insights that contribute to the broader and long-term use of foundation models. Moreover, our research provides new perspectives in understanding the adaptation of VLMs in terms of maintaining the representation alignment beyond the training distribution. We believe this insight can contribute to a deeper understanding of existing adaptation strategies (e.g., weight ensemble [42]) and the development of new methods. Besides, while our study focuses on video recognition, which has wide-ranging applications such as surveillance, it is crucial that concerns regarding privacy and individual rights are thoroughly considered prior to practical deployment.

B More Implementation Details

In Table 7, we present the hyper-parameters set for optimization. For both the Cross-dataset and Base-to-novel settings, we trained the model for 15 epochs. All experiments are conducted using 3 NVIDIA GeForce RTX 4090.

For cross-dataset evaluation, the methods are evaluated on three official splits or the full dataset of UCF-101 and HMDB-51. For Kinetics-600, we adopt the three splits provided by [4]. Each split contains 160 categories out of 220 new categories that do not exist in K400. We report the average Top-1 accuracy and the standard deviation on three splits. To further enhance the generalization performance, we incorporate the rephrased text descriptions from FROSTER [15] and the weight ensemble technique [42] into our method.

For base-to-novel evaluation, we do not apply weight ensemble since it may affect the model's performance on the base category. Following the previous work [15], we employed the rephrased text descriptions in this setting. Besides, we do not apply the rephrased text descriptions to the SSv2 dataset.

C More Ablations

Table 7: Hyper-parameter details during fine-tuning.

	Value
Optimization details	
Batch size	192
Optimizer	AdamW
Weight decay	0.2
Adam β_1,β_2	0.9, 0.999
Learning rate (Projection)	5e-5
Learning rate (CLIP layers)	3.75e-6
Learning rate decay	Cosine
Training epochs	15
Linear warm-up epochs	5 (cross-dataset), 2 (base-to-novel)
Augmentation	
RandomResizedCrop	
Area	[0.08, 1.00]
Aspect ratio	[3/4 , 4/3]
Crop size	224
Random Horizontal Flip	0.5
Random Gray scale	0.2

Training and efficiency analysis of TACO. We report the training and inference cost of different components in Table 8. Training time is measured on 3 NVIDIA GeForce RTX 4090 GPUs with a batch size of 192, and GPU-hours are computed from the wall-clock time. We also report GFLOPs, memory usage, parameter count, and throughput (TP) under the same ViT-B/16 setting. As shown in the table, the specialization projection introduces almost no extra overhead over the baseline, with nearly unchanged GFLOPs, memory, parameter count, and throughput. Adding $\mathcal{L}_{\text{RSD}}$ increases training cost and memory/FLOPs due to the extra teacher forward pass and anchor-based distillation, but the overall overhead remains modest. These results show that TACO improves performance with limited additional cost.

Table 8: Training and efficiency analysis of TACO.

Method	GPU-hours	GFLOPs	Mem (G)	Param (M)
Baseline	36.3	432	12.0	86.2
+ spec. proj.	36.4	432	12.1	86.2
+ $\mathcal{L}_{\text{RSD}}$	37.8	576	15.2	86.2

Effects of varying momentum coefficients α. We further study the impact of the teacher update mechanism and the momentum coefficient on generalization performance. As shown in Table 9, using a frozen teacher model (i.e., momentum $= 1.0$) yields inferior results, suggesting that an entirely static teacher may overly constrain the student and hinder the acquisition of new task-specific knowledge. In contrast, the EMA update strategy provides a better balance between preserving structural consistency and gradually incorporating newly learned knowledge. Among the tested values, $\alpha = 0.9998$ gives the best performance on both UCF and HMDB. Using either a larger or smaller momentum leads to slight degradation, indicating that an appropriate update rate is important for maintaining the consistency of the OOD space structure during adaptation.

Table 9: Effects of EMA update mechanism and varying momentum coefficients.

Momentum	UCF	HMDB
1.0	84.2	53.7
0.99998	84.2	54.0
0.9998	**84.9**	**54.2**
0.998	84.1	53.8

Comparison with conventional regularizations. We further compare our Specialization Projection with conventional regularization strategies, including dropout and weight decay. As shown in Table 10, both dropout and weight decay provide only limited improvements in generalization performance. In contrast, specialization projection consistently achieves better results on all three benchmarks, with clear gains on UCF, HMDB, and K600. This suggests that the benefit of our design does not mainly come from generic overfitting reduction. Rather, by explicitly decoupling the representation space from the optimization space, our strategy better prevents the shared visual representation from being directly overspecialized to the fine-tuning categories.

Table 10: Comparison with conventional regularization strategies.

Method	UCF	HMDB
Dropout	82.9	51.6
Weight decay	82.8	52.0
Spec. Proj.	**83.7**	**53.1**

Effects of adding the specialization projection during fine-tuning and evaluation. In this ablation study, we further investigate the effect of adding the specialization projection on the generalization of representations during training and evaluation. Adding SP during evaluation indicates that the projected representation is used. As shown in Table 11, decoupling the representation space facilitates enhancing model generalization. Besides, discarding the projection layer during evaluation is important, as the projected representation is more prone to overfitting the fine-tuning data distribution.

Table 11: Effects of adding the specialization projection during fine-tuning and evaluation.

Fine-tuning	Evaluation	UCF	HMDB
-	-	82.8	52.0
✓	-	**83.7**	**53.2**
✓	✓	83.3	52.2

Effectiveness of each proposed component on the fine-tuned model with the frozen text encoder. In Table 12, we study the effectiveness of each proposed component when freezing the text encoder during fine-tuning, which is a common design choice in VLMs adaptation. Our proposed components deliver significant improvements in generalization performance, demonstrating the versatility of our method under diverse fine-tuning scenarios.

Table 12: Effectiveness of each proposed component on the fine-tuned model with the frozen text encoder.

Spec. Proj.	$\mathcal{L}_{\text{RSD}}$	UCF_f	HMDB_f	K600_f
		81.5	50.4	69.7
✓		82.8	52.7	72.1
	✓	83.2	52.2	72.3
✓	✓	**83.8**	**53.6**	**73.1**

Effects of varying loss balance coefficients λ for relative structure distillation. In Table 13, we investigate the effects of varying loss balance coefficients λ in fine-tuning. As shown in the table, incorporating our relative structure distillation objective effectively enhances the generalization performance. A relatively small loss balance coefficient of 0.4 yields the best overall results.

Table 13: Effects of varying loss balance coefficients.

λ	UCF	HMDB
0.0	83.7	53.2
0.4	**84.9**	54.2
0.7	84.6	54.2
1.0	84.5	**54.4**

Effects of varying learning rate for specialization projection. We study the decoupling effect of different learning rates applied to projection parameters in Table 14. Specifically, we fixed the learning rate applied to the CLIP parameters and adjusted the learning rate for the projection layer. Results indicate that the learning rate ratio of around 10 offers the optimal results for space decoupling. Ratios below this threshold cannot effectively decouple the space, while higher ratios risk compromising the alignment between modalities.

Table 14: Effects of varying learning rate for specialization projection.

learning rate	UCF	HMDB
5×10^{-6}	83.1	52.8
5×10^{-5}	**83.7**	**53.2**
5×10^{-4}	82.6	52.3

D Theoretical Analysis of Relative Structure Distillation (RSD)

To better clarify the theoretical intuition behind RSD, we present a theoretical analysis to better support its underlying rationale.

D.1 OOD Geometric Non-identifiability of Standard KL Distillation

Standard KL distillation directly matches the teacher and student predictions on ID samples:

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}\left(\sigma\left(v_t W_t^\top/\tau\right)\middle|\sigma\left(v_s W_s^\top/\tau\right)\right).$$

Geometrically, each logit is the projection of a visual embedding onto an ID text direction. Let $\mathcal{U} = \text{rowspan}(W_t)$ be the subspace spanned by the ID text embeddings and $\text{rank}(W_t) < D$. Standard KL therefore observes only the component of a visual embedding within $\mathcal{U}$:

$$v_t W_t^\top = (v_t P_{\mathcal{U}}) W_t^\top.$$

where $P_{\mathcal{U}}$ denotes the orthogonal projector onto $\mathcal{U}$. Consequently, changes in the orthogonal complement $\mathcal{U}^\perp$ are invisible to the direct logits. To see this, keep the text embeddings fixed and consider a norm-preserving transformation of the visual space:

$$v_s = v_t T, \qquad W_s = W_t, \qquad T = P_{\mathcal{U}} + A P_{\mathcal{U}^\perp},$$

where A is any orthogonal transformation on $\mathcal{U}^\perp$. Since T leaves every ID text direction unchanged,

$$v_s W_s^\top = v_t T W_t^\top = v_t W_t^\top.$$

Thus, the visual geometry can change in $\mathcal{U}^{\perp}$ while all ID logits and the standard KL loss remain unchanged. However, an unseen text embedding may contain a component in $\mathcal{U}^{\perp}$, in which case

$$v_s w_{\text{OOD}}^{\top} \neq v_t w_{\text{OOD}}^{\top}$$

in general. Hence,

$$\boxed{\text{identical ID logits} \not\Rightarrow \text{identical OOD cross-modal structure.}}$$

The text branch admits a symmetric argument, and jointly updating both encoders introduces additional unconstrained degrees of freedom. This motivates geometric measurements beyond the directions covered by the ID texts.

D.2 Random Geometric Anchors as Isotropic Probes

Each anchor is sampled and normalized as

$$g_m \sim \mathcal{N}(0, I), \qquad z_m = \frac{g_m}{||g_m||_2}.$$

By rotational invariance, z_m is uniformly distributed on the unit hypersphere and satisfies

$$\mathbb{E}[z_m z_m^{\top}] = \frac{1}{D} I.$$

For any representation change $e \in \mathbb{R}^{1\times D}$,

$$\mathbb{E}_{z_m}\left[(e z_m)^2\right] = \frac{1}{D}||e||_2^2.$$

Therefore, random anchors have equal expected sensitivity to every direction and introduce no systematic bias toward the ID subspace or any particular semantic direction.

For M anchors, define

$$S_Z := Z^{\top} Z = \sum_{m=1}^{M} z_m z_m^{\top}, \qquad \mathbb{E}[S_Z] = \frac{M}{D} I.$$

A finite Z probes only finitely many directions in one iteration. However, Z is isotropic in expectation, and repeated resampling continually introduces new probing directions throughout training. It thus provides a data-independent mechanism for probing the shared CLIP representation space.

D.3 Additional Geometric Constraints Beyond Direct Logit Matching

Let $w_{a,n}$ denote the n-th row of W_a, where $a \in \{t, s\}$, and define the symmetric visual-text relation matrix

$$C_{a,n} := \text{sym}(v_a^{\top} w_{a,n}) = \frac{1}{2}\left(v_a^{\top} w_{a,n} + w_{a,n}^{\top} v_a\right).$$

A direct CLIP logit observes only its trace:

$$v_a w_{a,n}^{\top} = \text{tr}(C_{a,n}).$$

Because softmax is invariant to a common logit shift, consider the teacher-student discrepancy for a class pair (n, k):

$$\Delta_{nk} := (C_{s,n} - C_{s,k}) - (C_{t,n} - C_{t,k}).$$

Matching the direct logits constrains only

$$\text{tr}(\Delta_{nk}) = 0,$$

which does not imply $\Delta_{nk} = 0$. Thus, a realizable nonzero trace-free relation change can remain invisible to standard KL.

By contrast, the anchored logit is

$$v_a Z^{\top} Z w_{a,n}^{\top} = \sum_{m=1}^{M} (v_a z_m)(z_m^{\top} w_{a,n}^{\top}) = \langle S_Z, C_{a,n}\rangle_F.$$

RSD therefore measures the class-relation discrepancy as

$$\langle S_Z, \Delta_{nk}\rangle_F = \sum_{m=1}^{M} z_m^{\top} \Delta_{nk} z_m.$$

Writing

$$S_Z = \frac{M}{D} I + E_Z, \qquad \mathbb{E}[E_Z] = 0,$$

gives

$$\langle S_Z, \Delta_{nk} \rangle_F = \frac{M}{D} \operatorname{tr}(\Delta_{nk}) + \langle E_Z, \Delta_{nk} \rangle_F.$$

The first term is a scaled direct-logit measurement. The second term, induced by finite-sample anisotropic fluctuations, provides the additional geometric observation.

For $z \sim \mathrm{Unif}(\mathbb{S}^{D-1})$ and any symmetric Δ,

$$\mathbb{E}_z \left[(z^\top \Delta z)^2 \right] = \frac{2||\Delta||_F^2 + \operatorname{tr}(\Delta)^2}{D(D+2)}.$$

In particular, if a nonzero relation change is invisible to direct logits, i.e., $\operatorname{tr}(\Delta_{nk}) = 0$, then

$$\boxed{\mathbb{E}_Z \left[\langle Z^\top Z, \Delta_{nk} \rangle_F^2 \right] = \frac{2M}{D(D+2)} ||\Delta_{nk}||_F^2 > 0.}$$

Hence, every nonzero discrepancy in the direct-logit null space produces a positive expected squared anchor response. Rather than providing semantics, the random anchors Z act as directional probes that expose otherwise unobserved teacher–student geometric discrepancies.

D.4 Implication for Suppressing OOD Alignment Shift

The above analysis shows that standard KL leaves a geometric null space in which cross-modal relations can change without affecting ID logits, whereas resampled isotropic anchors make such changes observable with positive expected squared responses. Any anchor-measured logit-gap discrepancy changes the softmax distribution and is therefore penalized by RSD loss. Since OOD concepts share the pretrained CLIP space and may depend on directions unconstrained by ID logits, **RSD reduces the freedom for cross-modal distortion, thereby suppressing destructive drift of the pretrained cross-modal structure in OOD regions.**

E More Analysis and Discussions

Analysis of Random Geometric Anchors Overlapping with ID Space As discussed in the main text, while our proposed random geometric anchors could theoretically overlap with the in-distribution training distribution, this probability is negligible during fine-tuning. This is due to the fact that CLIP's representation space occupies a highly narrow conical region within the D-dimensional hypersphere. In Figure 7, we visualize the distribution of **maximum cosine similarities** between 10K randomly sampled geometric anchors and the CLIP training text representations. We observe that all similarity scores are below 0.25, indicating that no anchors overlap with the training distribution. This finding is consistent with the conclusions presented in [26]. Furthermore, we calculated the intrinsic dimension of the CLIP training text representations according to [8]. The resulting dimension is 11, which is significantly lower than the embedding dimension of 512. This further confirms that the ID space occupies only a minuscule fraction of the 512-hypersphere.

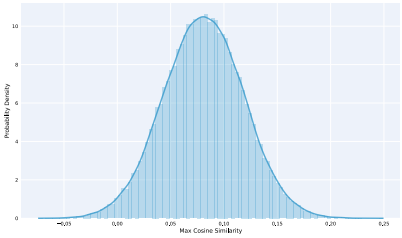


Figure 7: Max cosine similarities between random geometric anchors and ID text representations

Discussion on image-domain generalization. Although TACO is designed for open-vocabulary video adaptation, its benefit is not limited to the video domain. As shown in Table 15, TACO achieves the best performance on ImageNet (IN), ImageNet-V2, ImageNet-S, and ImageNet-R, while remaining competitive on IN-A. These results show that TACO improves not only in-domain accuracy but also robustness under distribution shift. This suggests that the gain of TACO does not mainly come from video-specific temporal modeling, but from preserving transferable cross-modal structure during adaptation. By preserving the relative structure of the pretrained CLIP space and reducing overspecialization to the source data, TACO remains effective beyond video recognition, indicating its potential as a general adaptation framework for vision-language models.

Table 15: Comparison with existing methods on image-domain generalization benchmarks.

Method	IN	IN-A	IN-V2	IN-S	IN-R
CoCoOp [52]	71.02	50.63	64.07	48.75	76.18
MaPLe [21]	70.72	50.90	64.07	49.15	76.98
PromptSRC [22]	71.27	50.90	64.35	49.55	77.80
WISE-FT [43]	72.38	51.07	65.29	49.72	72.48
MMA [47]	71.00	51.12	64.33	49.13	77.32
MMRL [12]	72.03	**51.20**	64.47	49.17	77.13
TACO	**73.18**	50.70	**65.55**	**50.52**	**78.20**

Discussion on the relation to LP-FT [24]. Compared with LP-FT, the key difference lies in the task setting and the resulting challenge. LP-FT studies single-modal transfer learning, where a pretrained vision encoder is adapted and evaluated under a fixed label space. In contrast, our work focuses on open-vocabulary adaptation in a CLIP-style joint visual-text embedding space, where the model is fine-tuned on known categories but evaluated on unseen ones. Under this setting, the main difficulty is not only preventing visual feature distortion, but also preserving cross-modal alignment in OOD space. Moreover, the motivation of LP-FT is rooted in the trade-off between linear probing and full fine-tuning for a single vision encoder, whereas our method is motivated by the inconsistency between fine-tuning and evaluation objectives in open-vocabulary adaptation. Based on this perspective, Section 2 further provides observations that are specific to our setting, including the source of adapted OOD generalization and the relationship between alignment shift and OOD performance. Correspondingly, our technical solution is also fundamentally different: rather than mitigating the coupled update between a classification head and a visual encoder, TACO explicitly regularizes the relative structure of the joint embedding space and decouples the representation space from the optimization space to better preserve cross-modal consistency during adaptation.

Difference from FROSTER. While FROSTER [15] also employs a CLIP teacher to preserve generalization, our method differs primarily in design motivation and supervision scope.

FROSTER is motivated by preventing student features from overfitting to the training distribution and therefore separately constrains the absolute visual and textual features to remain close to their teacher counterparts, without explicitly modeling the cross-modal alignment between them. In contrast, RSD directly regularizes geometric drift in the OOD space, treating cross-modal alignment shift—rather than modality-wise feature deviation—as the key property to preserve for generalization. In terms of supervision scope, FROSTER matches teacher and student features only on training samples, confining its supervision to the ID distribution. RSD instead explicitly extends supervision into the OOD space—a region FROSTER cannot reach by construction.

Difference from relation-based distillation. While conceptually similar to relation-based distillation methods (e.g., CRCD [53]), our method differs primarily in what relations are constructed and where they are anchored.

These methods aim to enhance discriminative organization among training categories by preserving the relative structure among training samples, matching sample-to-sample relations (e.g., via NCE-style objectives) computed entirely within the training batch—so their relational supervision remains confined to the ID distribution. In contrast, RSD constructs relations between ID embeddings and anchors explicitly placed beyond the ID manifold, allowing it to capture ID-to-OOD relational shifts, regardless of how the relations among training samples are defined. This distinction is empirically supported by Table 1c, supporting that the gain comes specifically from anchoring relations in a constructed OOD space, rather than from relation matching itself.

Metrics for structural preservation on unseen categories. For the comparison in Table 2, all three metrics use the original CLIP space as a reference. **Text-direction cosine** is the mean cosine similarity between corresponding pairwise unseen-class directions $w_i - w_j$ in the adapted model and CLIP, where w_i denotes the text embedding of class i. **Text kNN overlap@5** is the average proportion of each unseen class's five nearest text neighbors shared with CLIP. **Cross-modal Spearman correlation** is the mean per-video rank correlation between the adapted model's similarities to all unseen-class texts and the corresponding CLIP similarities. Higher values indicate better preservation of class directions, local text neighborhoods, and cross-modal rankings, respectively.

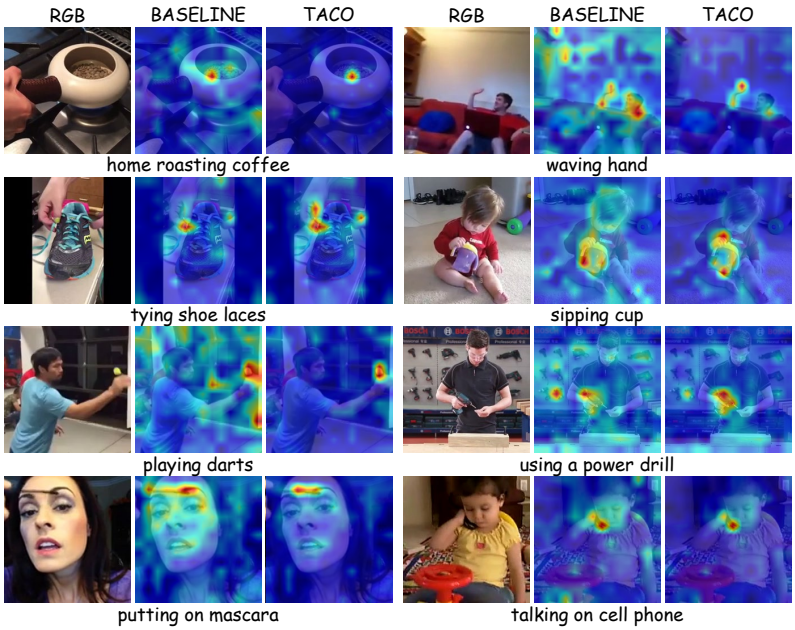


Figure 8: Visualization of model attention maps.

F Visualizations

Model Attention. To better understand what generalization knowledge the model captures when processing videos of unknown categories, we visualize the attention heatmap of the model in Figure 8. The attention maps are calculated between the text embedding and the patch embeddings output by the model. As shown in the figure, our model can focus more intently on key objects or actions, with less distraction from the background or other irrelevant objects. This phenomenon suggests that our model can learn more generalized video knowledge while maintaining the strong spatial understanding capabilities of the original CLIP during fine-tuning.

G Textual Prompts Used in Evaluation

Following the previous work [55], we adopt a set of hand-craft textual prompt templates to generate text embeddings during the evaluations. Following CLIP [37], we perform prompt ensembling over the 28 templates in order to provide comprehensive semantics, as listed in Table 16.

H Dataset Details

Kinetics-400 [20] is a large-scale dataset in the video domain. The dataset contains ∼240k training videos and ∼20k validation videos in 400 human action categories, with an average length of 10 seconds. The high quality of the dataset makes it the most popular benchmark for video recognition

Kinetics-600 [3] is an extension of Kinetics-400, consisting of ∼392k training videos, ∼30k validation videos, and ∼60k test videos in 600 human action categories. The dataset contains an additional 220 new action categories over Kinetics-400. We evaluate the zero-shot performance on 220 new categories and adopt three splits provided by the previous work [4]. We use its test set for evaluation and report the average performance on three splits.

UCF-101 [39] is an action recognition dataset that contains 13,320 videos in 101 action categories, collected from YouTube. There are three official splits of training data and validation data.

HMDB-51 [23] contains 7,000 videos in 51 action categories, collected from movie clips and web videos. There are three official splits of the dataset, each with 3,570 training data and 1,530 validation data samples.

Table 16: Textual prompt templates of TACO.

Templates
'a photo of {category}.'
'a photo of a person {category}.'
'a photo of a person using {category}.'
'a photo of a person doing {category}.'
'a photo of a person during {category}.'
'a photo of a person performing {category}.'
'a photo of a person practicing {category}.'
'a video of {category}.'
'a video of a person {category}.'
'a video of a person using {category}.'
'a video of a person doing {category}.'
'a video of a person during {category}.'
'a video of a person performing {category}.'
'a video of a person practicing {category}.'
'a example of {category}.'
'a example of a person {category}.'
'a example of a person using {category}.'
'a example of a person doing {category}.'
'a example of a person during {category}.'
'a example of a person performing {category}.'
'a example of a person practicing {category}.'
'a demonstration of {category}.'
'a demonstration of a person {category}.'
'a demonstration of a person using {category}.'
'a demonstration of a person doing {category}.'
'a demonstration of a person during {category}.'
'a demonstration of a person performing {category}.'
'a demonstration of a person practicing {category}.'

Something-Something V2 [10] is a temporal-heavy dataset that requires the fine-grained temporal understanding capability of the model. It contains 220,000 videos in 174 action categories.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction summarize the TACO framework, its two components, and its empirical scope. These claims are supported by the analyses in Sec. 2, the method in Sec. 3, the experiments in Sec. 4, and the geometric analysis in Appendix D.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly discusses limitations in Appendix A, including weaker gains on temporal-heavy datasets and sensitivity to the EMA momentum parameter.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Appendix D provides the geometric assumptions and derivations underlying RSD, including the ID-logit null space, the isotropy of random anchors, and the expected squared response to class-relation discrepancies. The analysis concerns geometric constraints and does not claim a universal guarantee of improved OOD accuracy.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The method is described in Sec. 3, and the evaluation protocols, hyperparameters, prompts, and datasets are specified in Sec. 4 and Appendices B, G, and H. The public implementation at `https://github.com/ZMHH-H/TACO` provides additional configuration and execution details.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The implementation is publicly available at https://github.com/ZMHH-H/TACO, with environment requirements, dataset-access links, configurations, and training and evaluation commands for the cross-dataset experiments. The benchmark datasets are available from their original providers.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 and Appendices B, G, and H describe the training and evaluation protocols, data splits, optimizer, learning rates, augmentations, epochs, prompt templates, and datasets.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 4 reports the main cross-dataset results as mean $\pm$ standard deviation across three evaluation splits, as described in Appendix B. These deviations measure variation across dataset splits, rather than variation across training random seeds.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Appendices B and C report the use of three NVIDIA GeForce RTX 4090 GPUs and the GPU-hours, GFLOPs, and memory usage of the main ViT-B/16 variants in Table 8. The paper does not provide a complete per-experiment compute breakdown or the total compute used by the full research project.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses standard benchmark datasets and pretrained models, does not involve human subjects, and includes discussion of broader impact and privacy-related concerns.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix A discusses positive impact for foundation model adaptation and negative concerns such as surveillance, privacy, and individual rights.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk generative models or new scraped datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The paper properly cites existing datasets and pretrained models, but it does not explicitly list the licenses or terms of use of these assets in the manuscript.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released TACO implementation at `https://github.com/ZMHH-H/TACO` is accompanied by a README, environment requirements, data-preparation instructions, training and evaluation commands, configuration files, and an Apache-2.0 license. The paper does not introduce a new dataset.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not part of the core method; LLM-generated categories are only used in an ablation as an alternative anchor source, not as an important methodological component.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.